NIEKOOPERATYWNE ZACHOWANIA SĘDZIEGO W TEŚCIE TURINGA

ANALIZA LOGÓW Z KONKURSU LOEBNERA

Aleksandra Kwiecień

Seminarium Gry w badaniach i edukacji

24 stycznia 2018

O KONKURSIE LOEBNERA

- pierwsza edycja odbyła się w **1990r.** z inicjatywy Hugh Loebnera
- konkurs opiera się na założeniach testu Turinga; w każdej rundzie sędzia rozmawia jednocześnie z botem oraz człowiekiem, a po zakończonej rozmowie podejmuje decyzję, kto jest kim i ocenia poziom programu, przyznając mu punkty
- brak ograniczeń dot. tematu rozmów, ograniczenie czasowe: 25 min na jedną rundę

- dla programu, który przejdzie test Turinga podczas konkursu (zostanie uznany za człowieka), jest przewidziana nagroda w postaci 100 000 dolarów oraz złotego medalu
- do tej pory żadnemu programowi nie udało się tego dokonać, ale najlepszy program w każdej edycji (wybierany na podstawie ocen sędziów) jest nagradzany 3000 dolarów i brązowym medalem
- ostatnie 2 edycje (2016, 2017) wygrał program **Mitsuku**, stworzony przez Steve'a Worswicka

KONKURS LOEBNERA 2017

- odbył się 16 września 2017r. w Bletchley Park, UK
- skład sędziowski:

Joanne Pransky: pierwsza na świecie psychiatra robotów
David Boyle: autor książki „Alan Turing: Unlocking the Enigma”
Jow Hewitt: strateg marki, Landor
Tom Cheshire: korespondent w dziedzinie technologii dla Sky News

- sędziowie przyznawali kolejno 4, 3, 2, 1 punktów (4 dla najbardziej „człowieczego” bota)

1. **Mitsuku** (13 points)
2. **Midge** (12 points)
3. **Uberbot** (8 points)
4. **Rose** (7 points)

IDEA NIEKOOPERATYWNOŚCI WG PLÜSSA

- Brian Plüss w swoich badaniach zajmował się politycznymi debatami
- wykorzystywał anotowanie dialogów w poszukiwaniu **NCF** (non-cooperative features), czyli niekooperatywnych cech wypowiedzi
- zaproponował miarę poziomu niekooperatywności **DNC** (degree of non-cooperation)

DNC = liczba NCF/liczba wypowiedzi w dialogu

TYPY ZACHOWAŃ NIEKOOPERATYWNYCH WG PLÜSSA

- **O** overlap;
- **GF** grounding failure;
- **UC** unsolicited comment;
- **I** interruption;
- **TC** topic change.

(Plüss, 2010)

H:	Well, the important aspect of this which it's very clear to bear in mind. . .	**GF**
P:	(interrupting) I'm sorry, I'm going to be frightfully rude but. . .	**I**
H:	Yes, you can. . .	
P:	(overlapping) I'm sorry. . .	**O**
H:	(overlapping) . . . you can put the question and I will give you, I will give you an answer.	**O**

H – Michael Howard (były minister spraw wewnętrznych Wielkiej Brytanii); P – Jeremy Paxman, dziennikarz BBC

(Plüss, 2010)

NIEKOOPERATYWNOŚĆ W KONKURSIE LOEBNERA

- w artykule „ Non-cooperative Strategies of Players in the Loebner Contest” (Łupkowski, Rybacka 2016) autorzy przedstawili pomysł na wykorzystanie NCF w analizie logów z konkursu Loebnera
- LC jest podobny do debat politycznych pod względem niektórych podstawowych założeń („rywalizacja” pomiędzy rozmówcami – dziennikarz/sędzia zadaje niewygodne pytania, których polityk/program chce unikać;
  w obu przypadkach zaleca się powstrzymywanie od wygłaszania opinii na temat przeciwnika)
- z powodu różnic pomiędzy LC i debatami politycznymi (brak rozmowy „twarzą w twarz”) dla konkursu Loebnera należało utworzyć inny zestaw NCF

NCF W KONKURSIE LOEBNERA

- **IR** wypowiedź nierelewantna
- **TC** zmiana tematu, która następuje po relewantnej wypowiedzi innego uczestnika
- **RD** odmowa udzielenia odpowiedzi/spełnienia prośby
- **PO** sędzia wyraża opinie na temat gracza
- **LP** długie pauzy, które mają wpływ na przebieg dialogu
- **IN** przeszkadzanie

(Łupkowski, Rybacka, 2016)

- **IR** (a judge or a participant fail to provide a relevant contribution)

J: What was the only thing you could think of?

P: *I can think of more than one only thing me could think of.*

- **TC** (topic change)

J: Tell me an event that you think is important.

P: *I don't know how to explain it to you.*

J: Describe the content of the board next to you.

- **RD** (request drop)

J: Would you ming describing the room we are in.

P: *I am not so sure if I would like to ming describing the room we are in.*

- **PO** (a judge expresses opinion about the player's performance)

J: How old are you then?

P: *I am 1 year, 4 months, 4 weeks, 2 days old.*

J: You are giving away the game here, don't you think?

- **LP** (*long pauses*)

P: *do you have a facebook account?*

J: [pause]

P: *are you married?*

- **IN** (*interrupting*)

J: Slow down, I can't keep up with you.

J: .

J: .

J: .

PROCEDURA ANOTACJI

1. Ustalenie zestawu NCF dla konkursu Loebnera.
2. Anotacja wypowiedzi przy użyciu zestawu NCF (przez 5 anotatorów).
3. Aby wypowiedź była uznana za niekooperatywną, potrzeba była zgodność przynajmniej 3 anotatorów, którzy ją za taką uznali.
4. Ustalenie poziomu zgodności anotacji (Kappa Fleissa); ewentualne eliminowanie tagów.
5. Ustalenie kategorii NCF każdej z wypowiedzi (NCF najczęściej wskazywane przez anotatorów)
6. Obliczenie DNC.

J: Would you like a glass of milk? **TC**

P: It doesn't sound fun to me.

J: Do you enjoy drinking milk?

P: I like Michael Douglas. **RD**

J: Have you drunk Michael Douglas? **IR**

P: I don't think I ever have drunk Michael Douglas. What's it like?

J – sędzia; P – uczestnik (program)

WYNIKI DNC

Program	DNC	Score	NCF Structure
		2009	
Levy	0.17	4.5	IR (56%); RD (40%)
Embar	0.36	5.5	RD (50%); IR (33%)
		2010	
Wilcox	0.42	2.5	RD (50%); IR (33%)
Medeksza	0.45	3.25	RD (50%); IR (33%)
		2011	
Wilcox	0.26	1.5	IR (47%); RD (15%)
Tutor	0.16	3.25	IR (65%); RD (29%)
		2012	
Chip	0.27	1.25	IR (53%); RD (14%)
Linguo	0.76	4	TC (73%); RD (22%)

Table 3 DNC measures and the most frequent NCFs for the *participants* of the Loebner contest editions 2009–2012

Judge	DNC	NCF Structure
	2009	
Round: Levy	0.06	TC (50%); PO (50%)
Round: Embar	0.15	RD (42.86%); PO (28.57%)
	2010	
Round: Wilcox	0.07	TC (53.85%); RD (23.08%)
Round: Medeksza	0.06	TC (83.33%); RD (8.33%)
	2011	
Round: Wilcox	0.10	IN (32.58%); RD (18.42%)
Round: Tutor	0.10	PO (42.86%); TC (28.57%)
	2012	
Round: Chip	0.06	TC (45.45%); RD (18.18%)
Round: Linguo	0.07	RD (58.33%); TC (41.67%)

Table 4 DNC measures and the most frequent NCFs for the *judges* of the Loebner contest editions 2009–2012

WNIOSKI

- miara DNC była niższa u sędziów niż botów (sędziowie byli bardziej kooperatywni)
- najczęstszym błędem popełnianym przez sędziów była zmiana tematu (TC), a przez programy – nierelewantne wypowiedzi (IR)
- bardzo ważny w LC jest dobór sędziów – aby utrudnić zadanie programom, należałoby wybierać na sędziów językoznawców i psychologów
- dużymi wadami konkursu są warunki, w jakich się odbywa – są to typowo „testowe", nienaturalne warunki do rozmowy; sędzia wie, że rozmawia albo z programem albo z człowiekiem i tylko wybiera jedną z opcji
- niekooperatywność w codziennych rozmowach może wpływać pozytywnie na pragmatyczny aspekt konwersacji

PLAN PRACY

1. Transkrypcja logów z 2017 do postaci dialogowej.
2. Wykonanie anotacji NCF.
3. Sprawdzenie jakości anotacji.
4. Ustalenie DNC dla logów z 2017.
5. Porównanie (Łupkowski, Rybacka 2016) z DNC dla LC 2017.

BIBLIOGRAFIA

- Łupkowski P., Rybacka A. (2016): Non-cooperative Strategies of Players in the Loebner Contest.
- Plüss, B. (2010). Non-Cooperation in Dialogue.